Towards a Universal Sentiment Classifier in Multiple Languages

Kui Xu & Xiaojun Wan

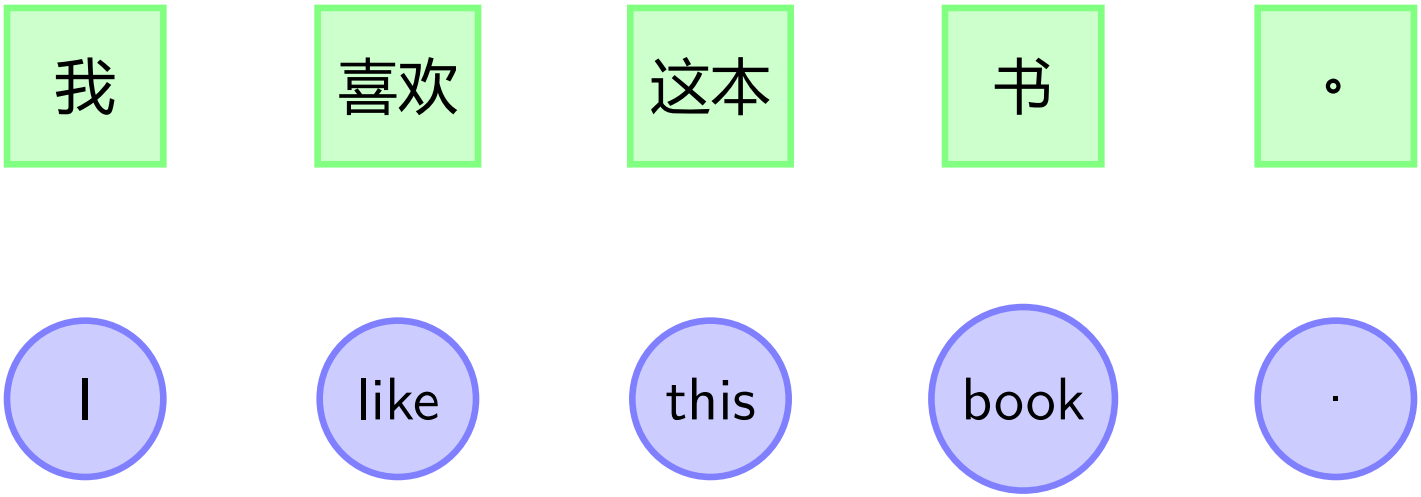
Peking University

2017.08.16

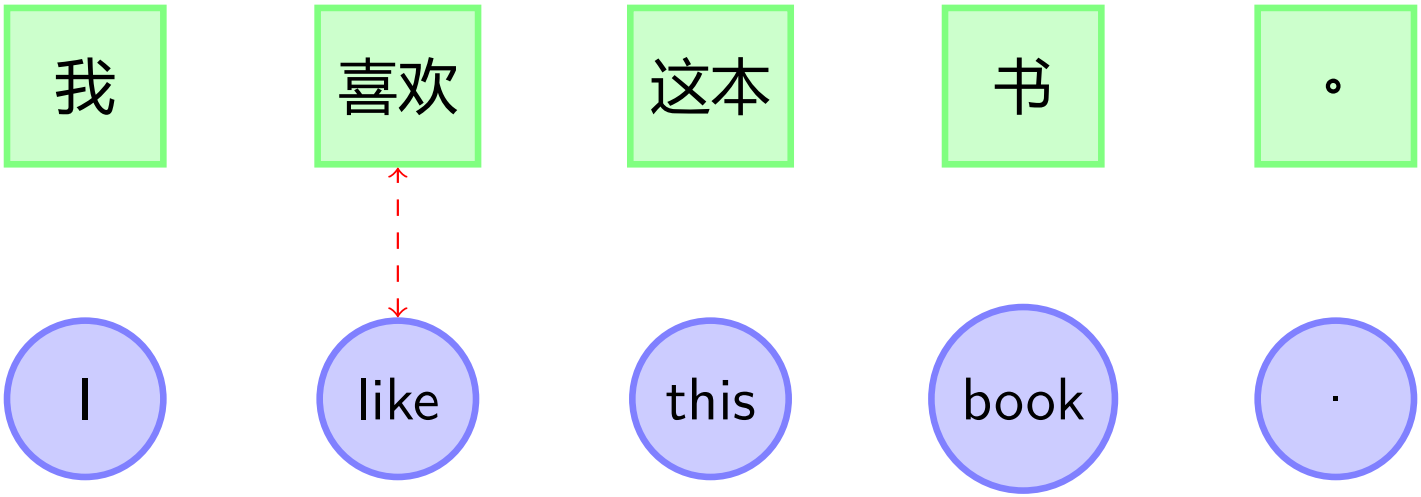
Abstract

- Existing sentiment classifiers usually work for only one specific language, and different classification models are used in different languages.
- We propose to learn multilingual sentiment-aware word embeddings simultaneously based only on the labeled reviews in English and unlabeled parallel data available in a few language pairs.

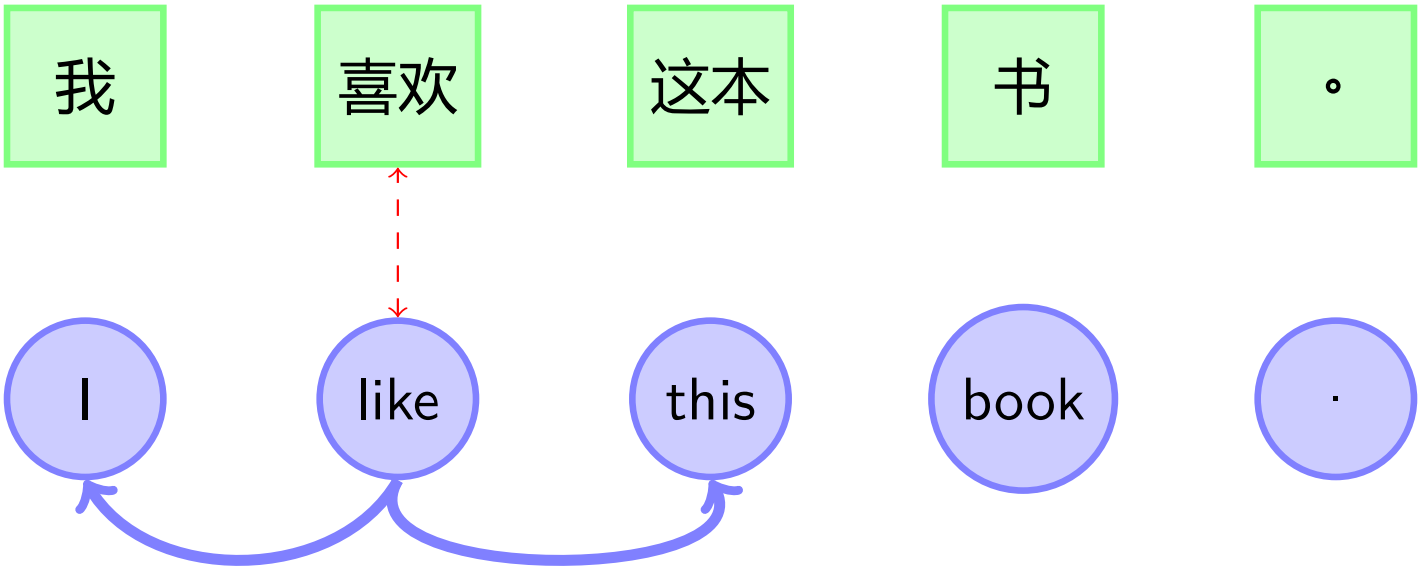
Bilingual word embeddings (BiSkip)



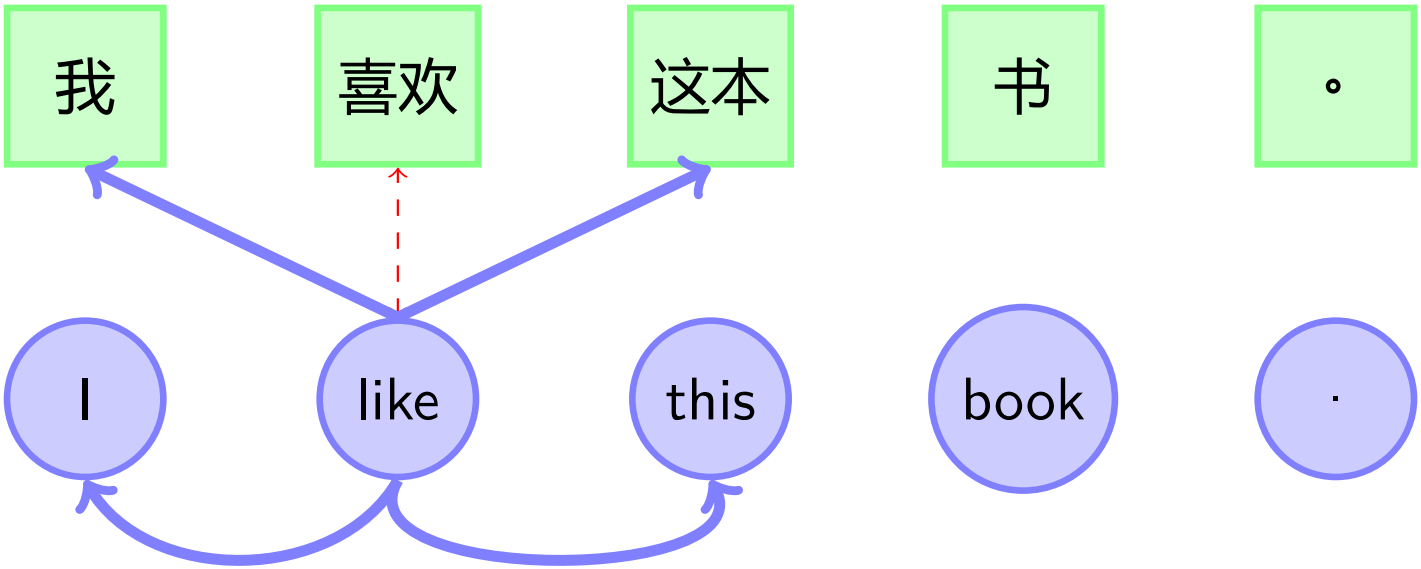
Bilingual word embeddings (BiSkip)



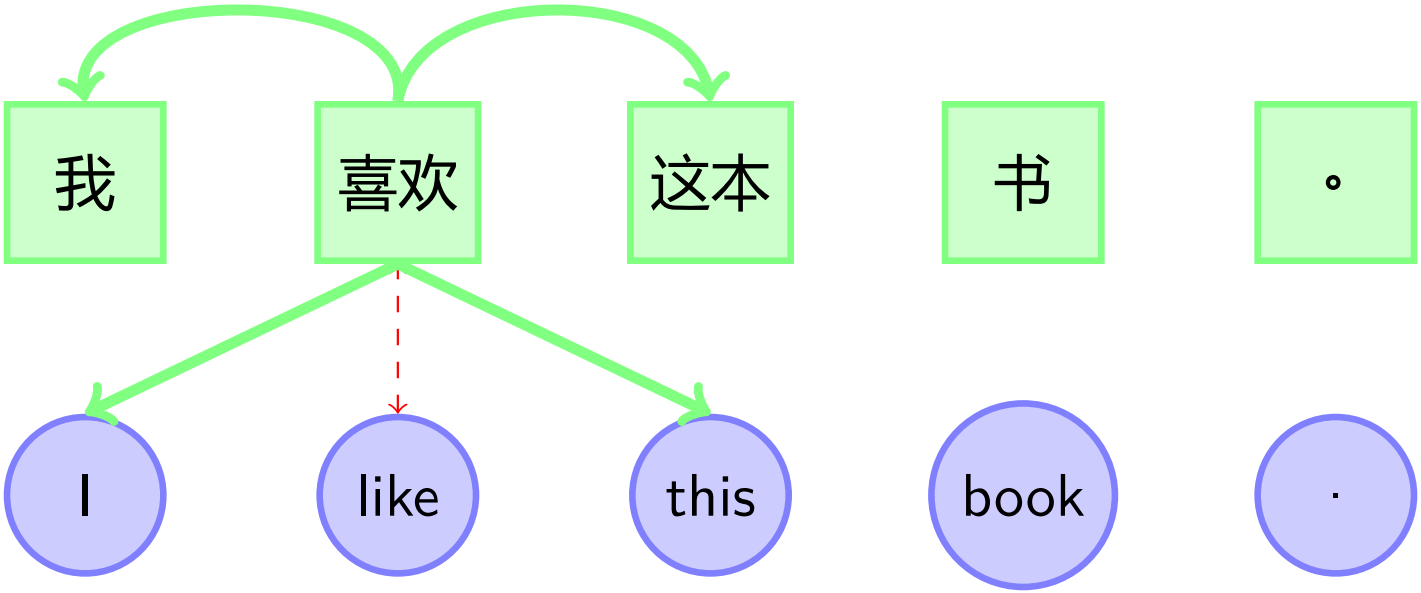
Bilingual word embeddings (BiSkip)



Bilingual word embeddings (BiSkip)



Bilingual word embeddings (BiSkip)



Bilingual word embeddings (BiSkip)

Formally, we assume a source language S with $|S|$ words and a target language T with $|T|$ words. We use s and t to represent a word from S and T, respectively. Given the bilingual parallel corpus $\mathcal{C}$ between language S and T, it can be divided into a corpus $\mathcal{C}_S$ in language S and a corpus $\mathcal{C}_T$ in language T. And we use a notation $S - T$ to indicate a parallel corpus between languages S and T.

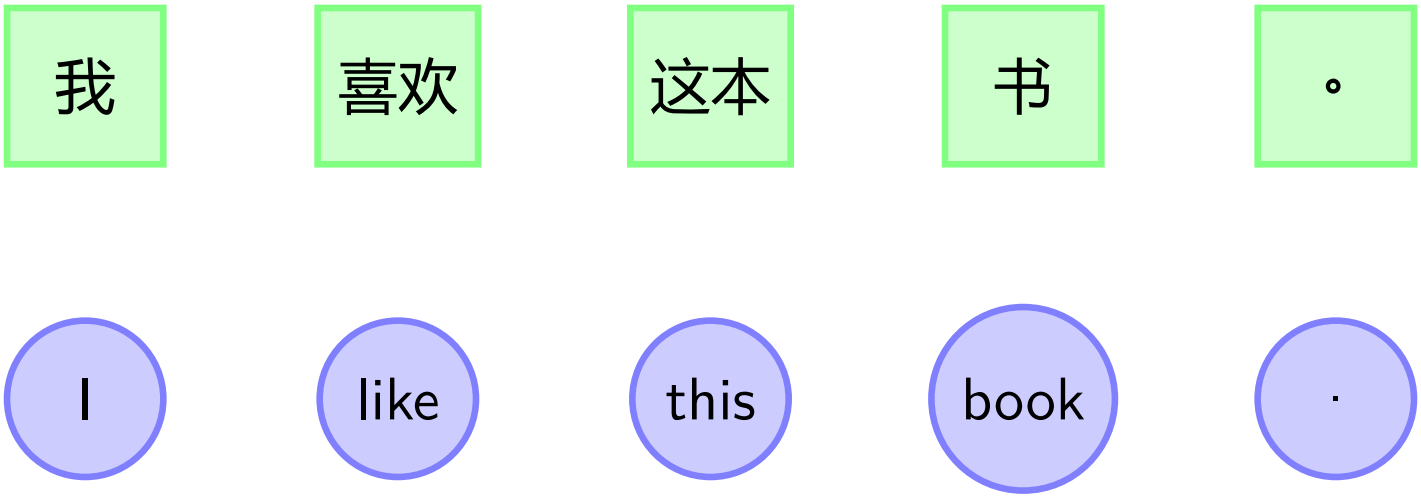
For corpus $\mathcal{C}_\mathcal{S}$, the monolingual constraint on itself ($\mathcal{C}_\mathcal{S} \rightarrow \mathcal{C}_\mathcal{S}$) is:

$$Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{S}) = \sum_{s \in \mathcal{C}_\mathcal{S}} \sum_{w \in adj(s)} \log p(w|s), \tag{1}$$

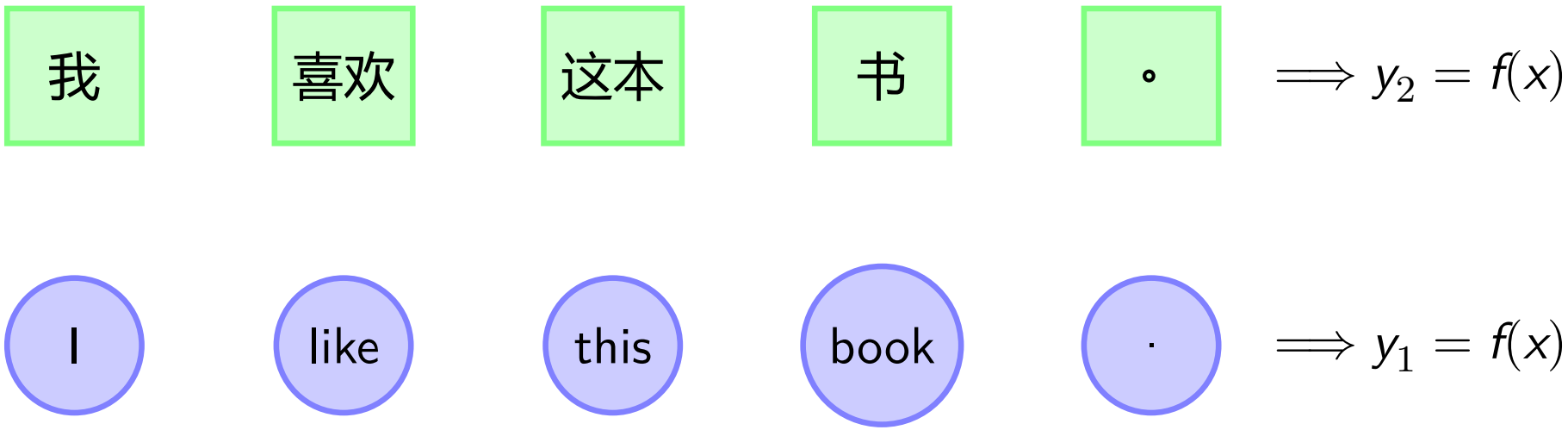
and the cross-lingual constraint on $\mathcal{C}_\mathcal{T}$ ($\mathcal{C}_\mathcal{S} \rightarrow \mathcal{C}_\mathcal{T}$) is:

$$Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{S}) = \sum_{s \in \mathcal{C}_\mathcal{S}} \sum_{w \in adj(t), s \leftrightarrow t} \log p(w|s) \tag{2}$$

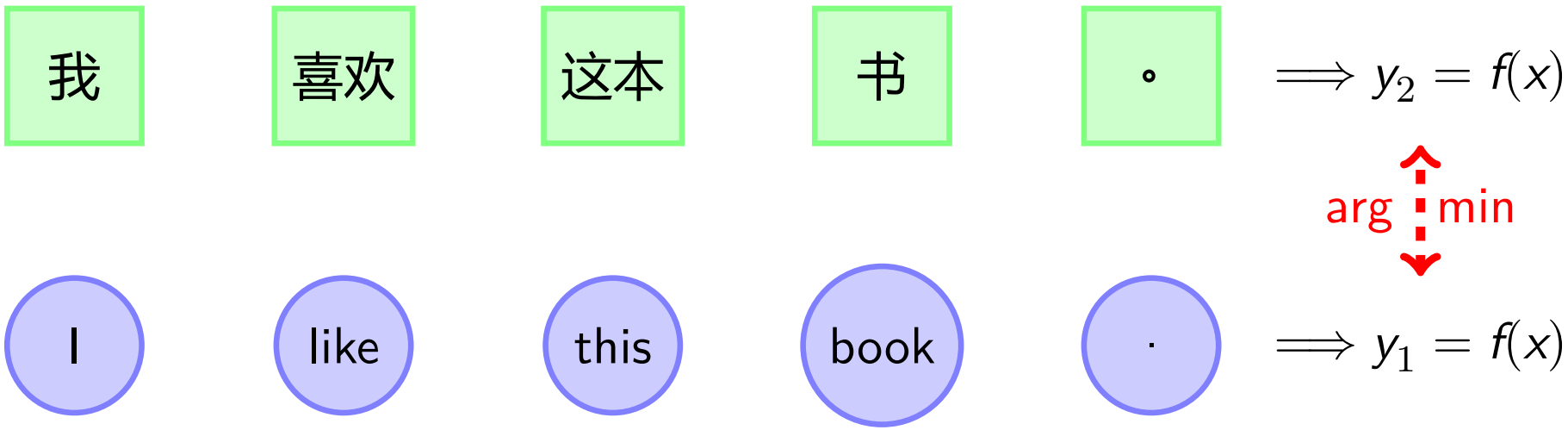
Bilingual word embeddings (BiCVM)



Bilingual word embeddings (BiCVM)



Bilingual word embeddings (BiCVM)

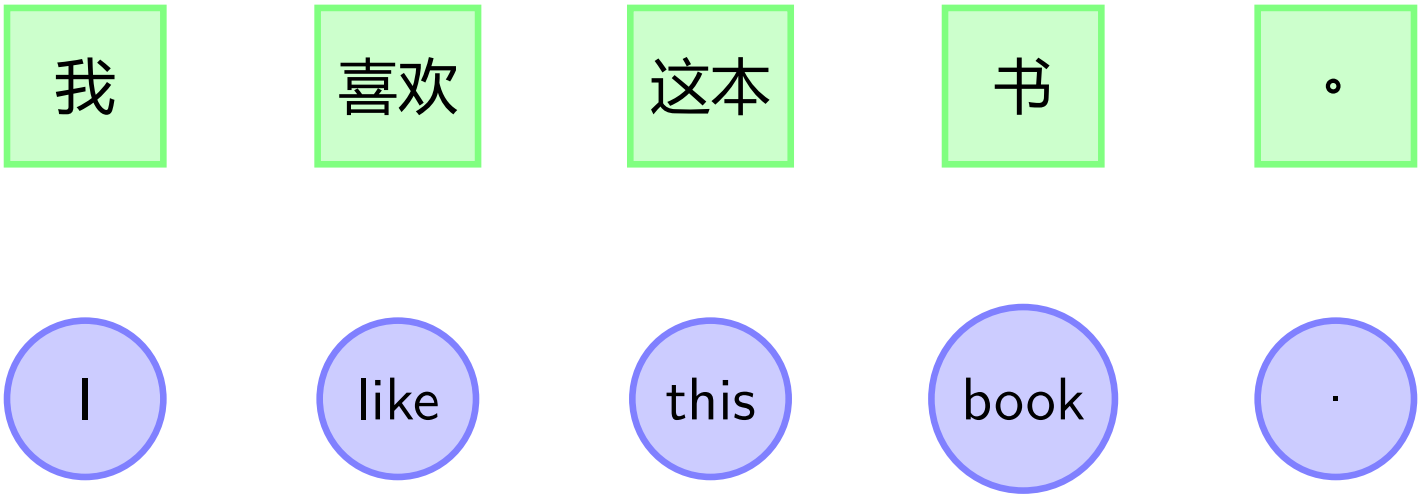


The sentiment classifier

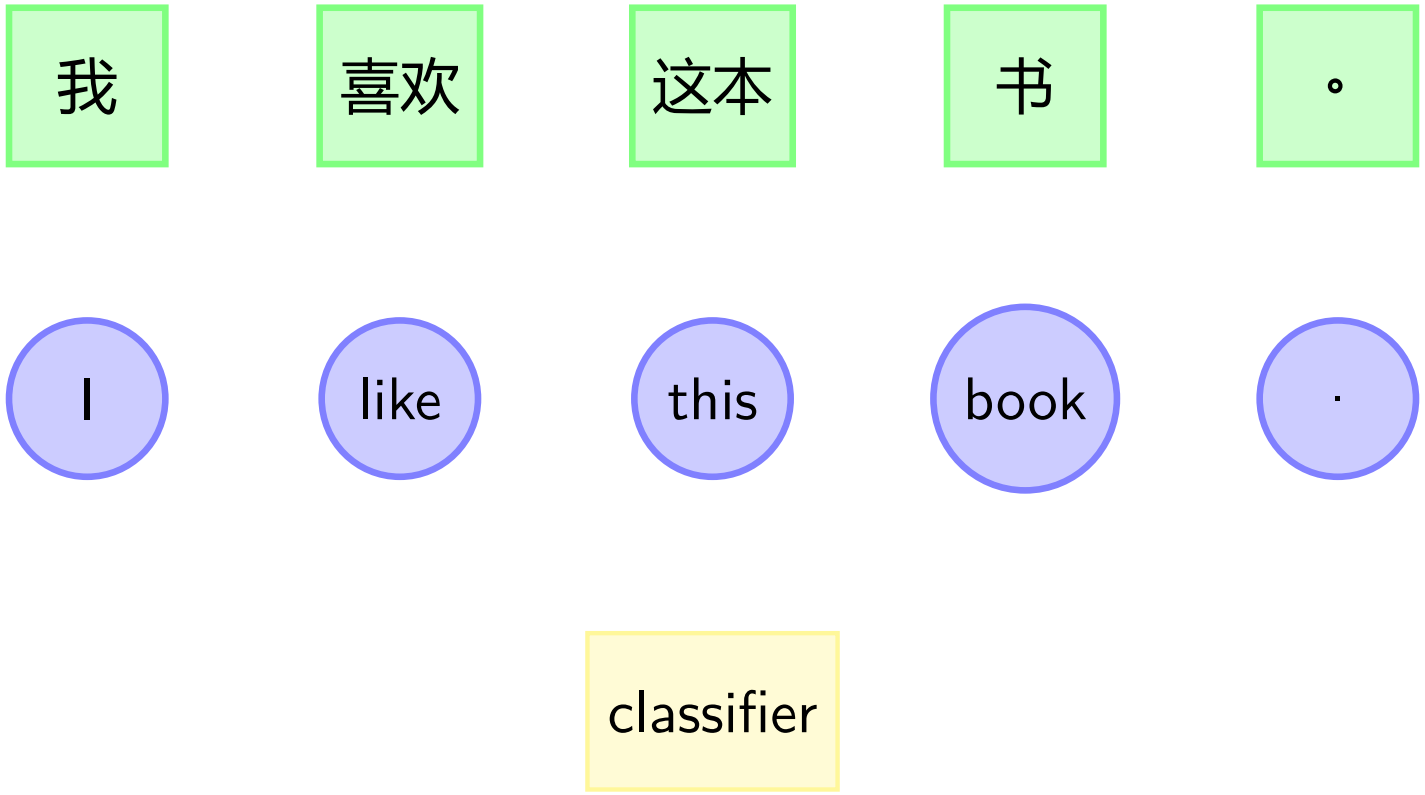
Given a labeled sentimental corpus $\mathcal{C}_{\mathcal{L}}$, we use S^* to represent a sentence in $\mathcal{C}_{\mathcal{L}}$ and w as a word in S^*. And x^T is a sum of word embeddings in S^*. We simply adopt the logistic regression classifier to enforce the sentiment constraint, and thus make the bilingual word embeddings absorb the corresponding sentiment information. The objective function is:

$$L(\mathcal{C}_{\mathcal{L}}) = \sum_{S^* \in \mathcal{C}_{\mathcal{L}}} y \log \sigma(Wx^T + b) + (1 - y) \log \sigma(1 - (Wx^T + b)) \qquad (3)$$

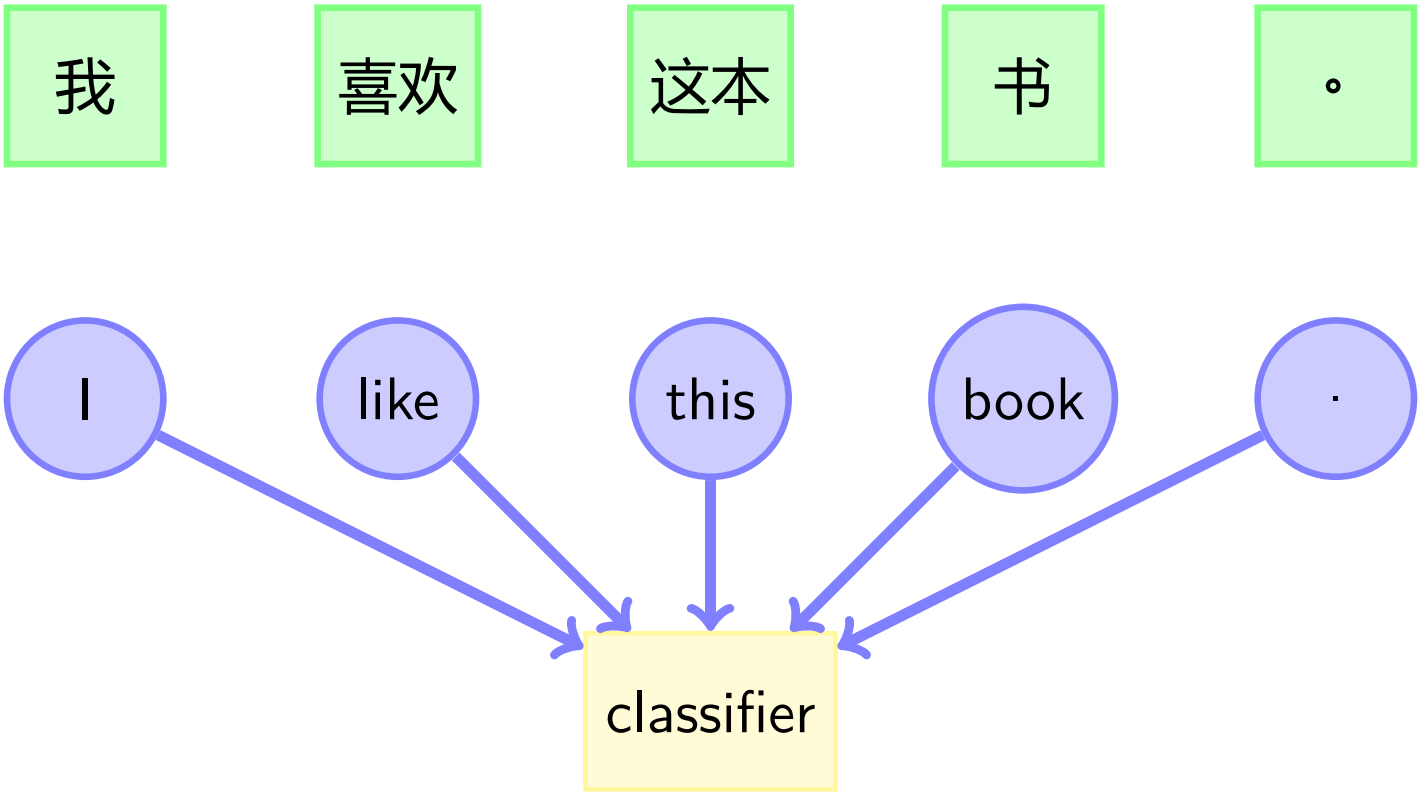
Bilingual Model (BM)



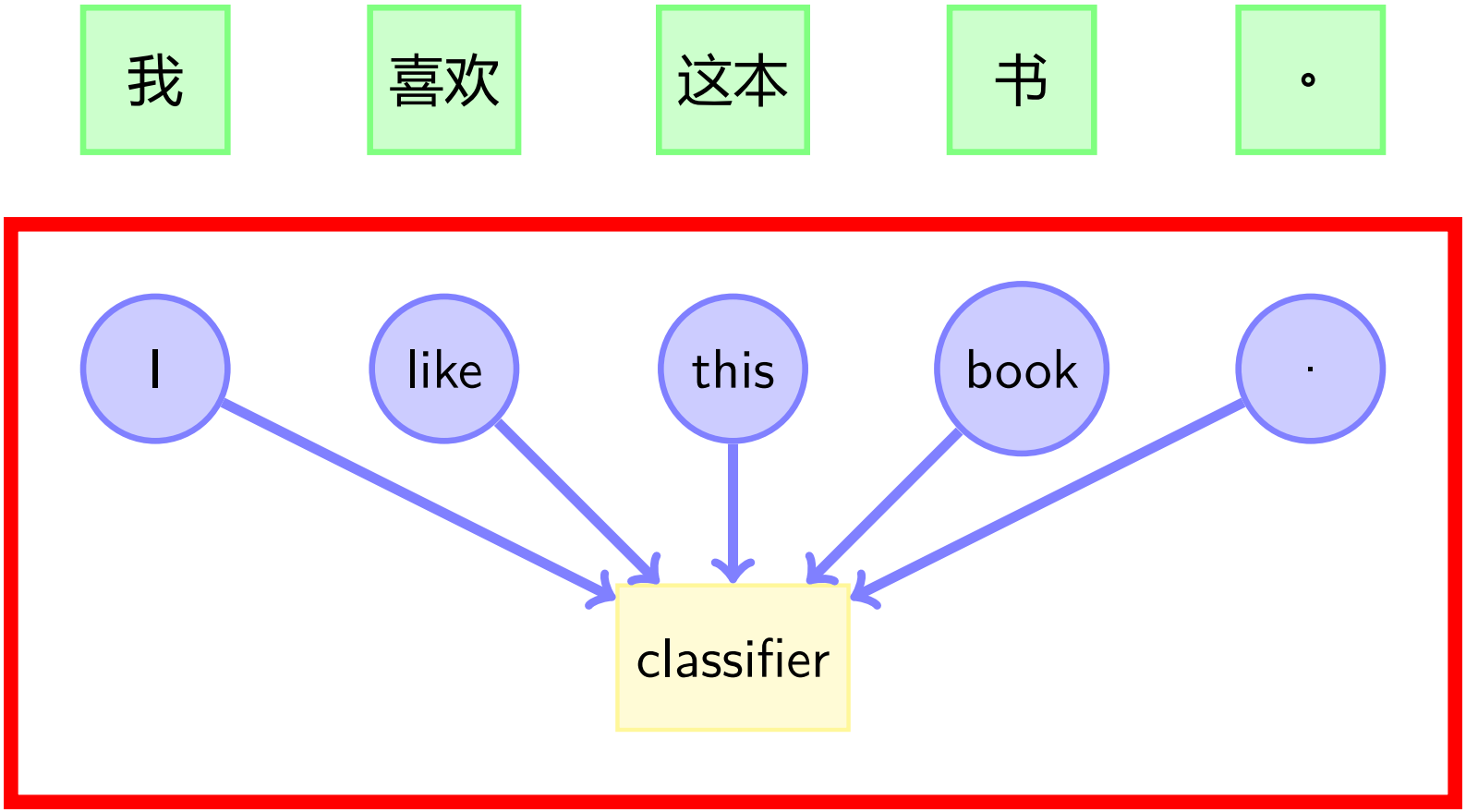
Bilingual Model (BM)



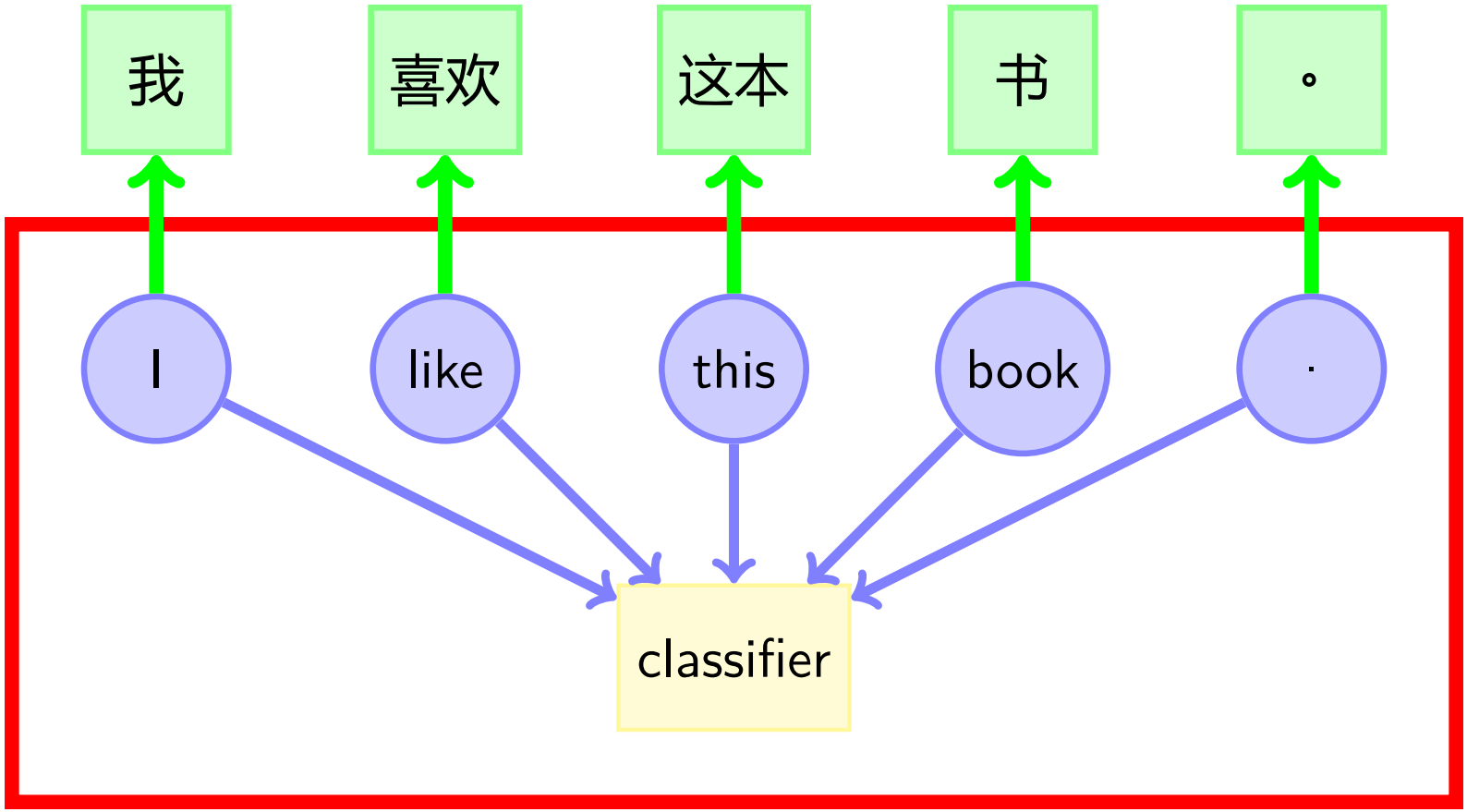
Bilingual Model (BM)



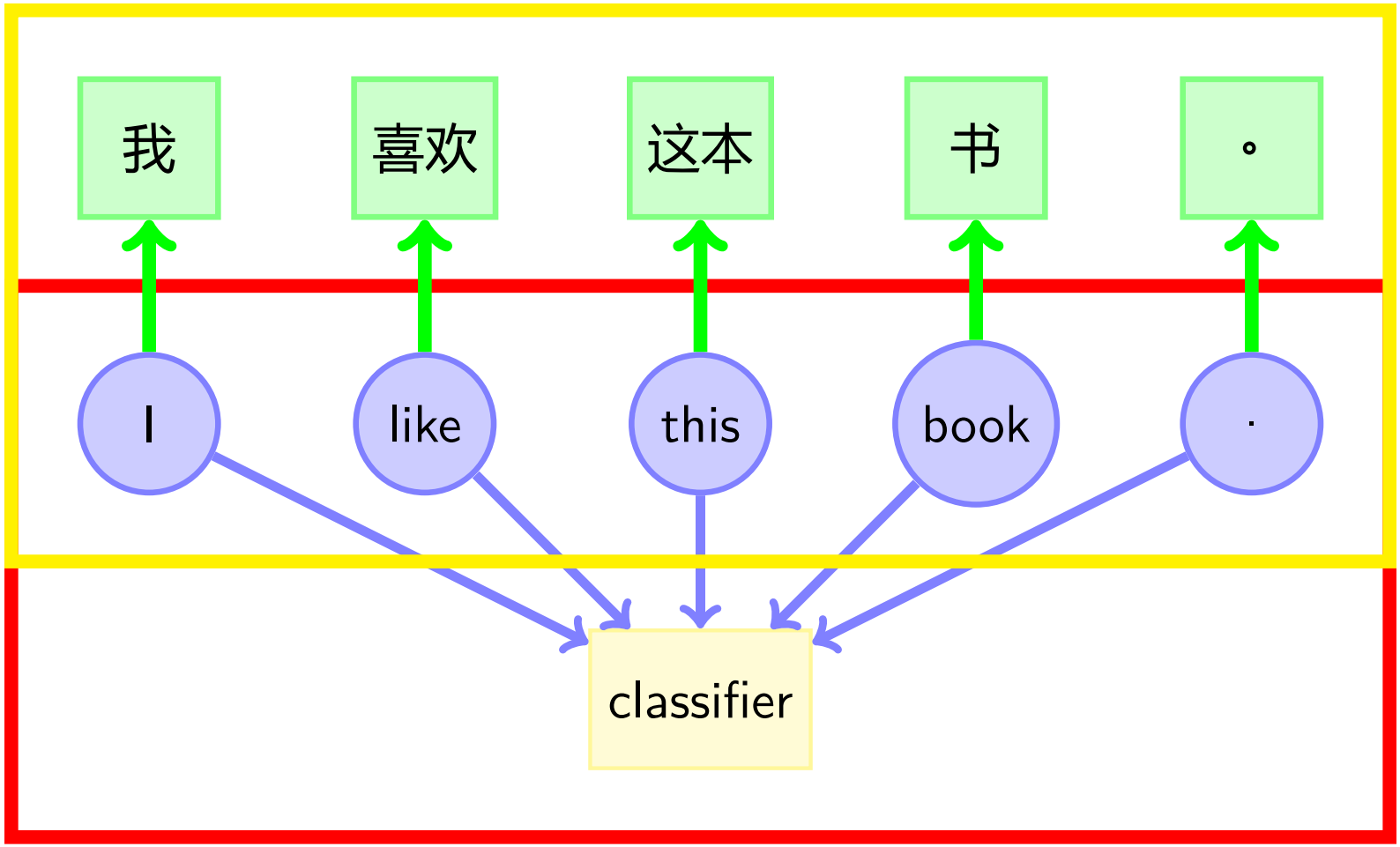
Bilingual Model (BM)



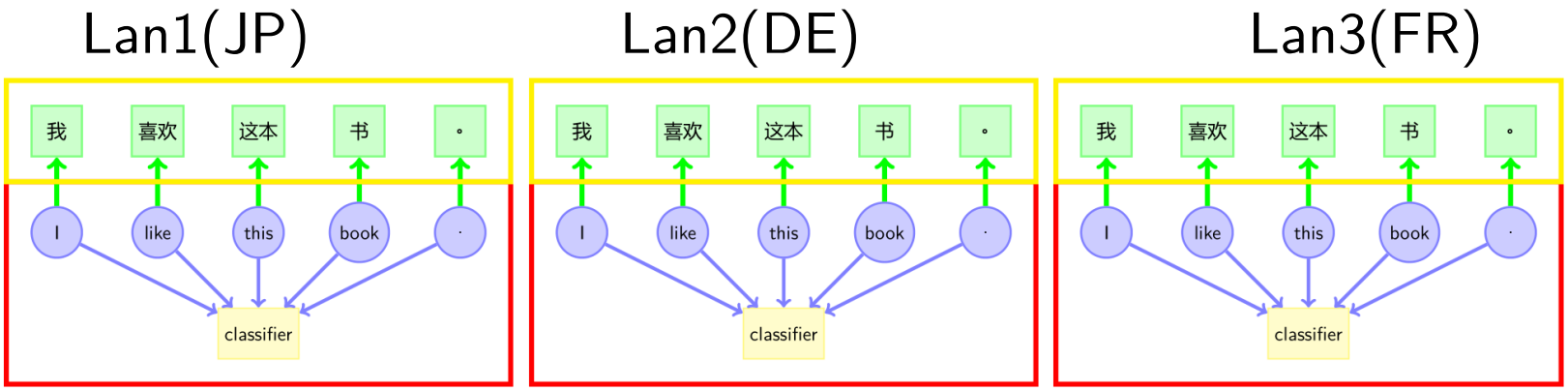
Bilingual Model (BM)



Bilingual Model (BM)

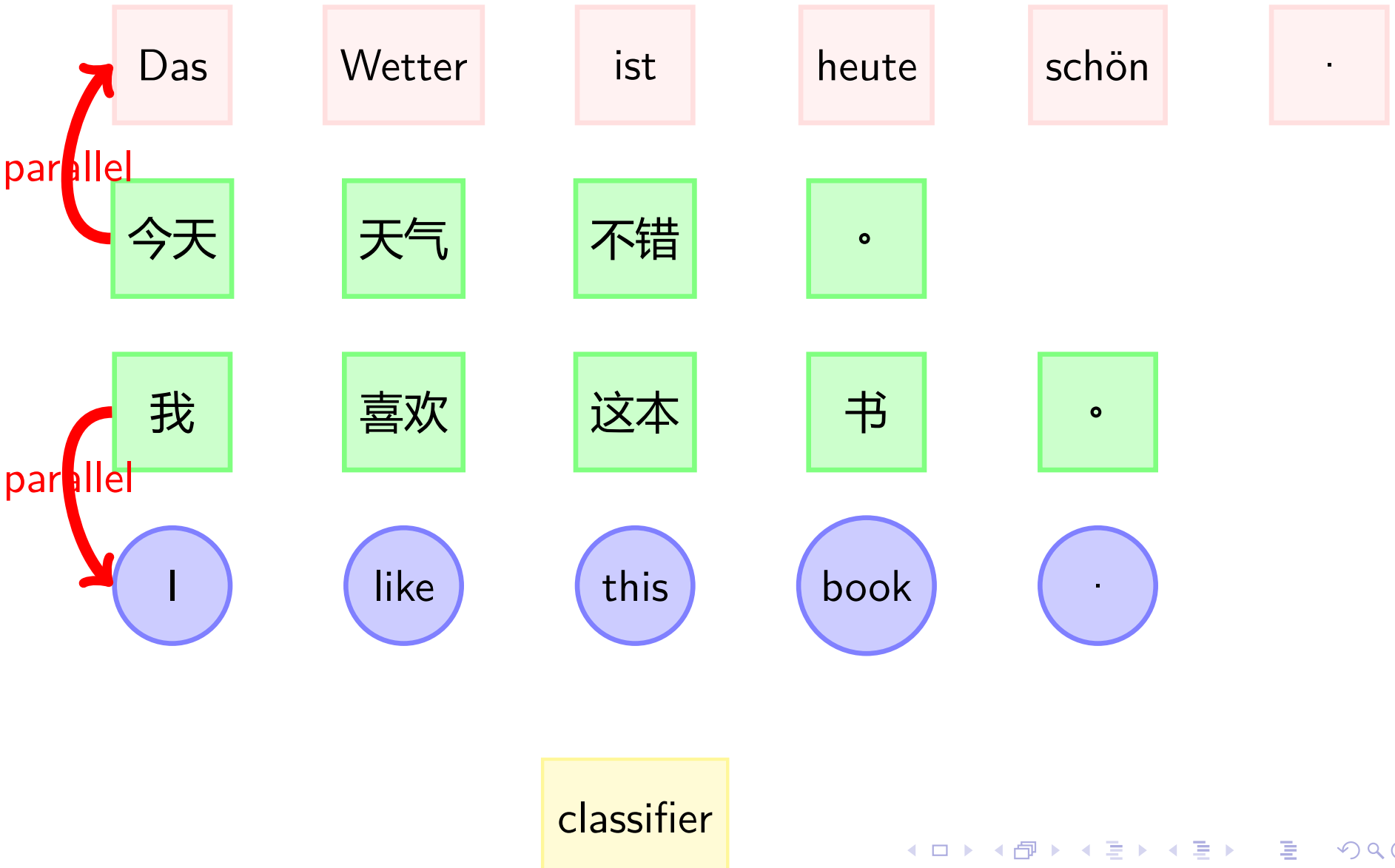


Bilingual Model (BM)

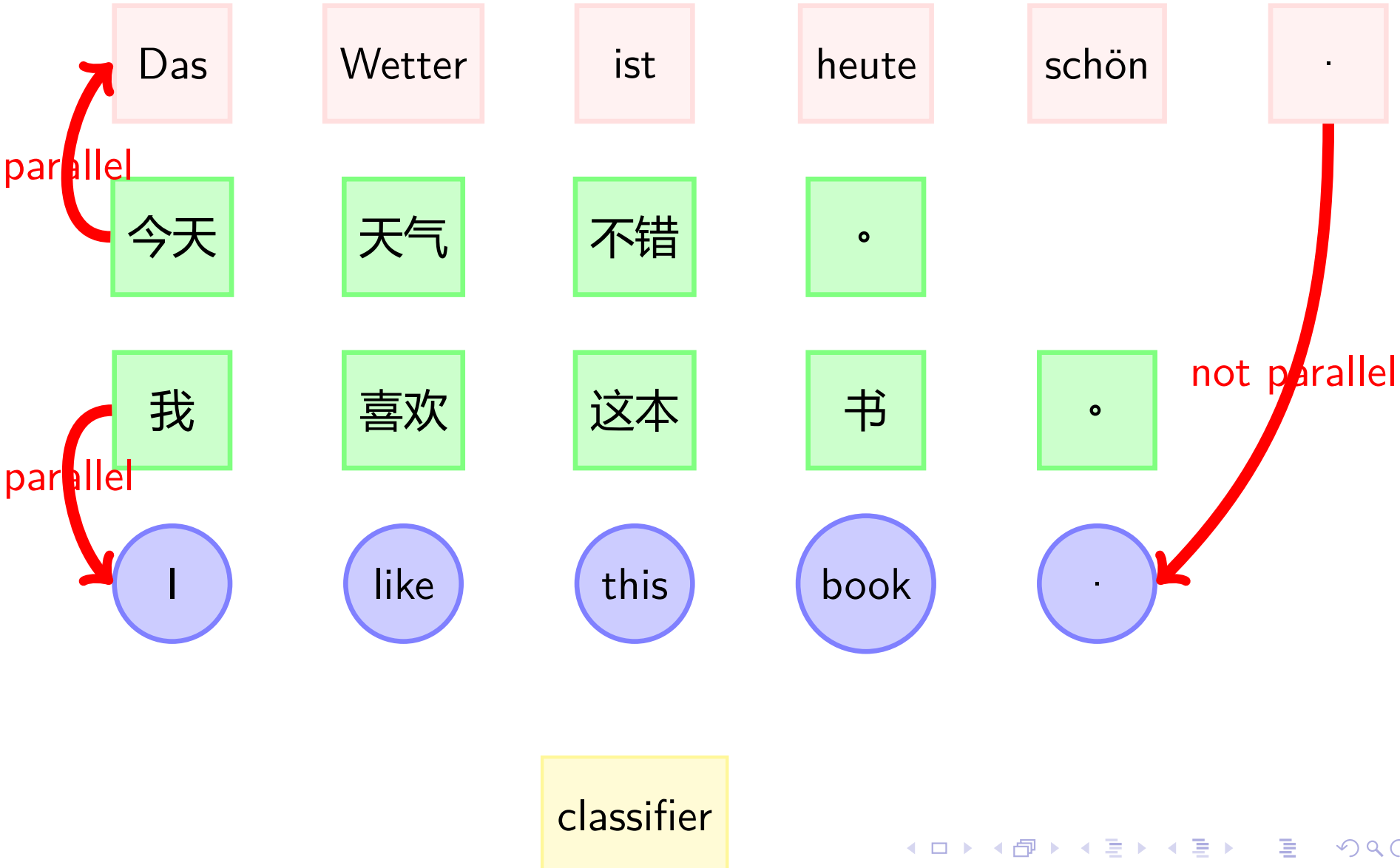


classifier

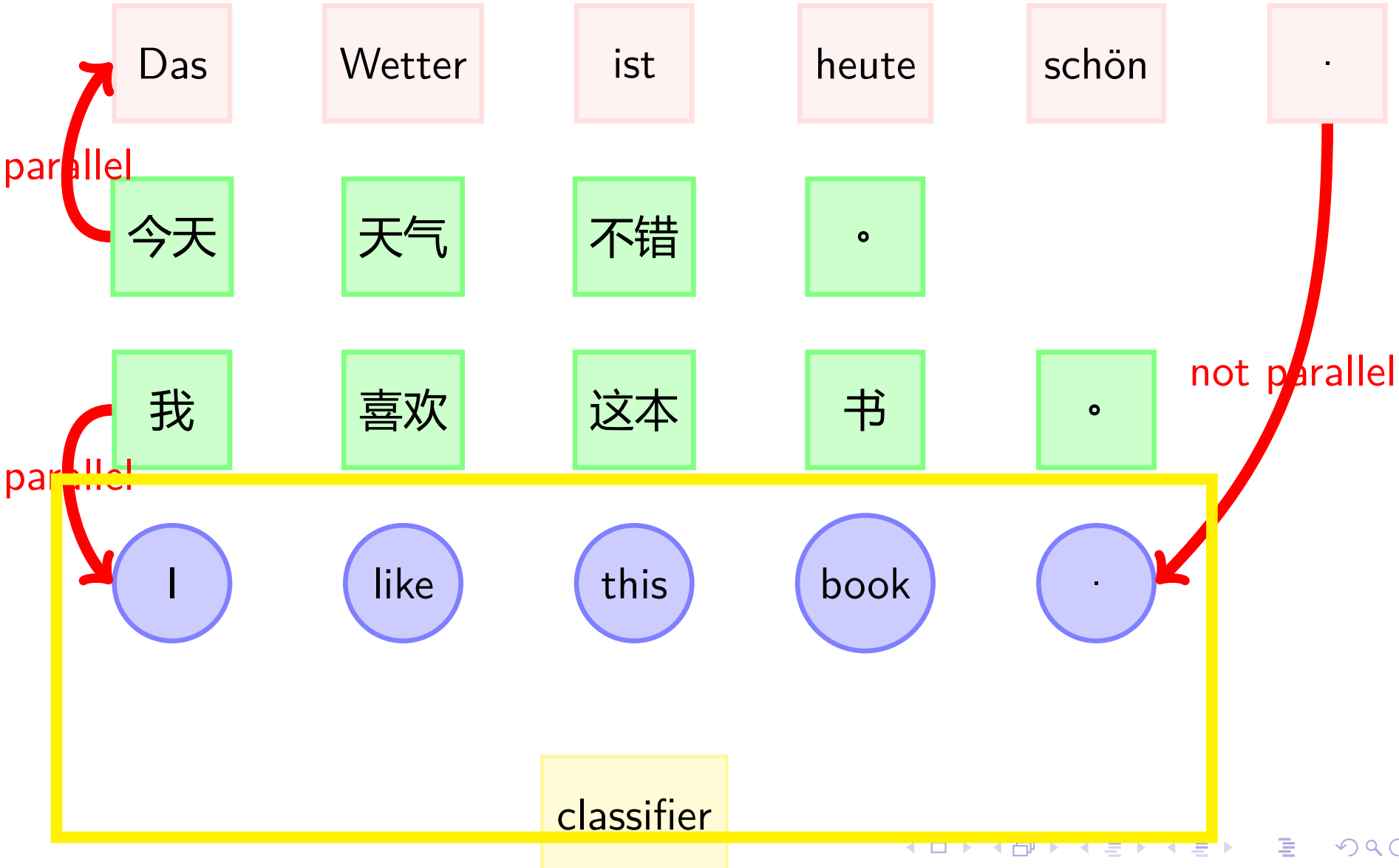
Pivot-Driven Bilingual Model (PDBM)



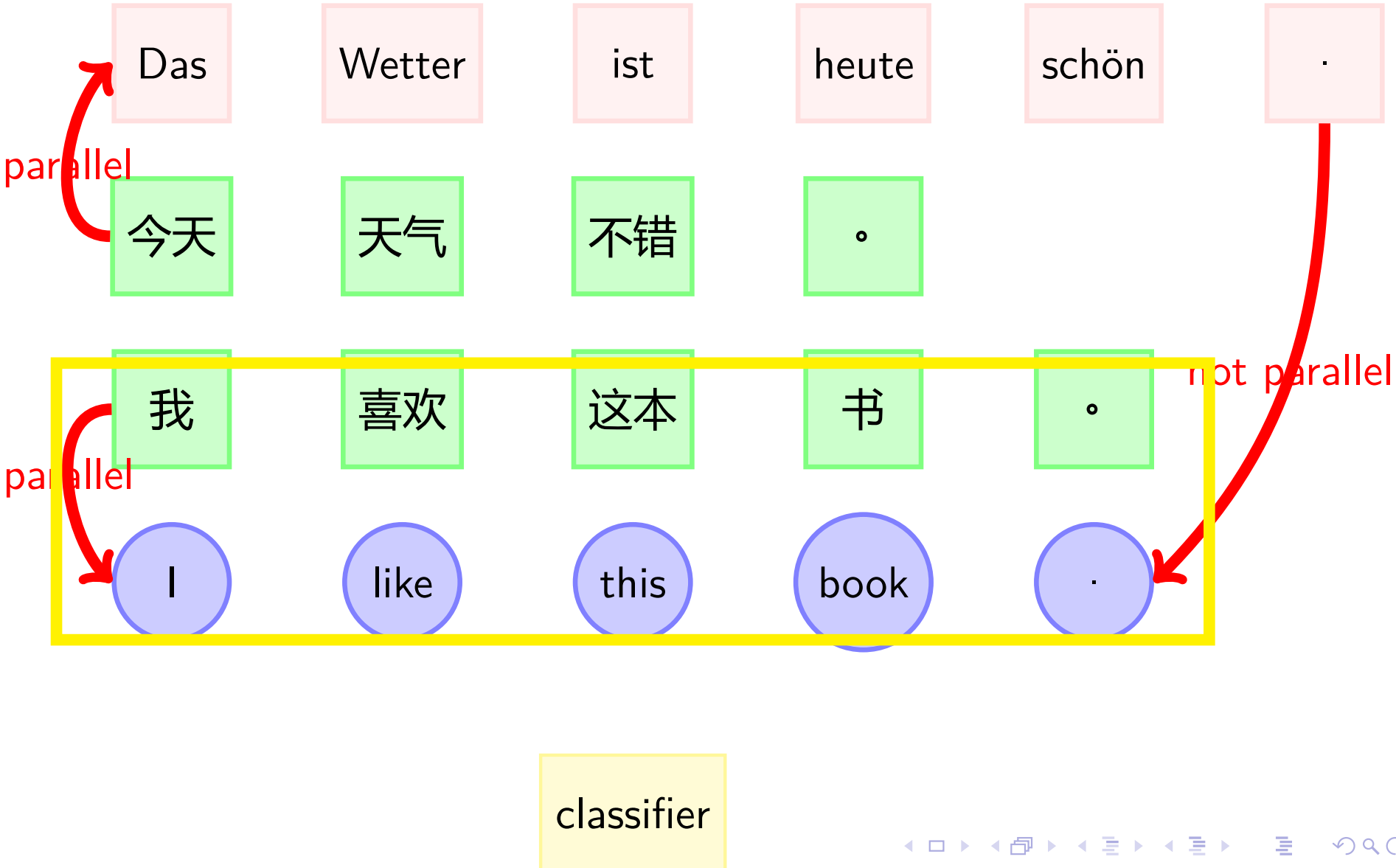
Pivot-Driven Bilingual Model (PDBM)



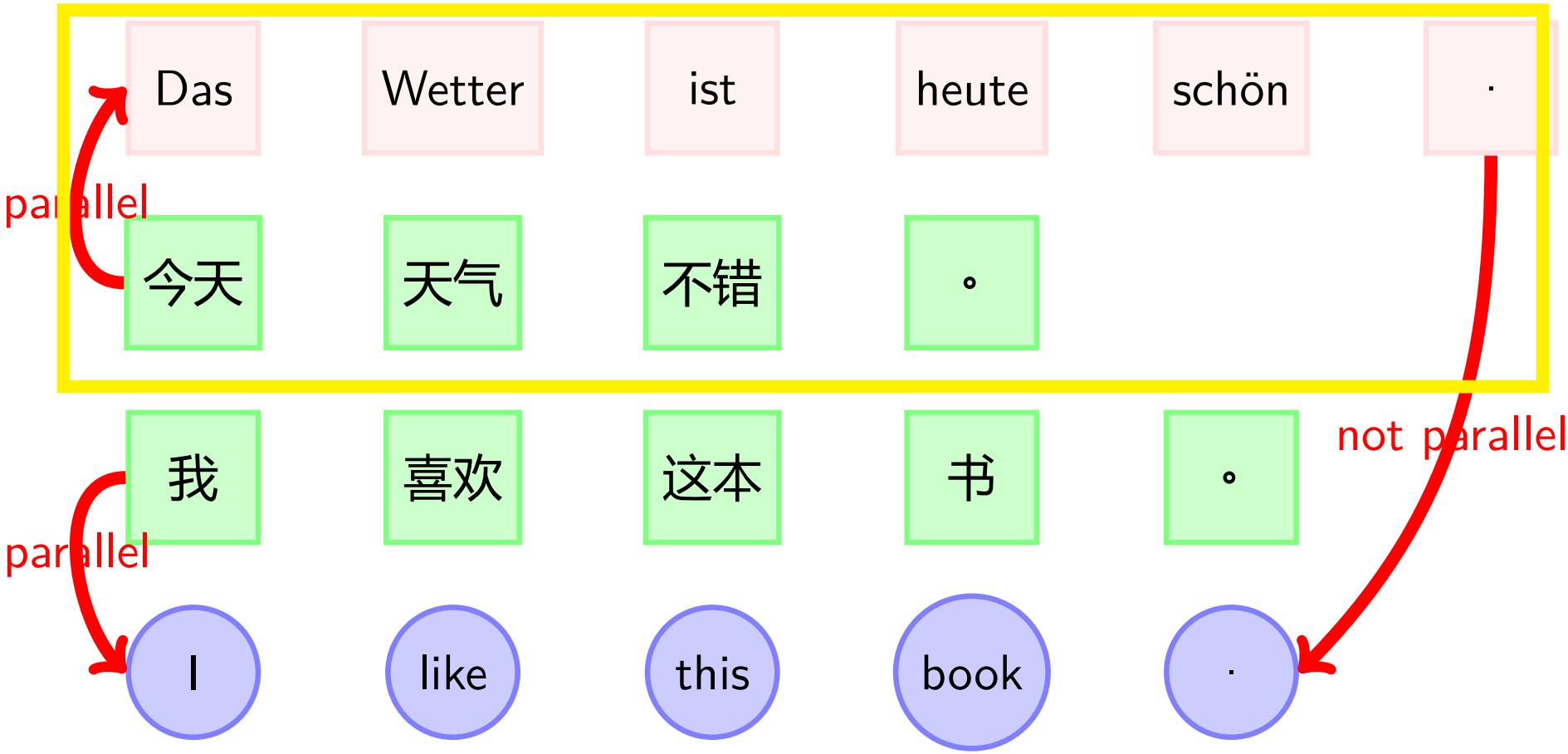
Pivot-Driven Bilingual Model (PDBM)



Pivot-Driven Bilingual Model (PDBM)



Pivot-Driven Bilingual Model (PDBM)



Universal Multilingual Model (UMM)

Even when a single pivot language cannot be found for languages T_i and S, we still can find two or more pivot languages $\{P_1, P_2, ..., P_M\}$ to form a pivot chain and the sentiment information in the source language can be passed through the pivot chain $(S - P_1 - ... - P_M - T_i)$ to the target language.

Results

TL	Domain	BM	PDBM	UMM	MT-BOW	CL-SCL	BSE	C
DE	book	**82.46**	81.97	81.65	79.68	79.50	80.27	7
	DVD	81.47	**82.67**	81.27	77.92	76.92	77.16	7
	music	**82.95**	81.93	81.32	77.22	77.79	77.98	7
FR	book	82.47	81.01	80.27	80.76	78.49	-	7
	DVD	**81.86**	81.68	80.27	78.83	78.80	-	7
	music	**81.51**	80.03	79.41	75.78	77.92	-	7
JP	book	70.93	71.59	71.23	70.22	**73.09**	70.75	7
	DVD	74.62	72.82	72.55	71.30	71.07	74.96	7
	music	**76.48**	76.26	75.38	72.02	75.11	77.06	7

Table: Comparison results (accuracy) on DE (German), FR (French) and JP (Japanese).

References

- Luong T, Pham H, Manning C D. Bilingual Word Representations with Monolingual Quality in Mind[C]// The Workshop on Vector Space Modeling for Natural Language Processing. 2015:151-159.
- Zhou H, Chen L, Shi F, et al. Learning Bilingual Sentiment Word Embeddings for Cross-language Sentiment Classification[C]// Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing. 2015.

Thank you!!

Q&A